

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-345239

(43)Date of publication of application : 14.12.1999

(51)Int.Cl. G06F 17/30

(21)Application number : 10-153231 (71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

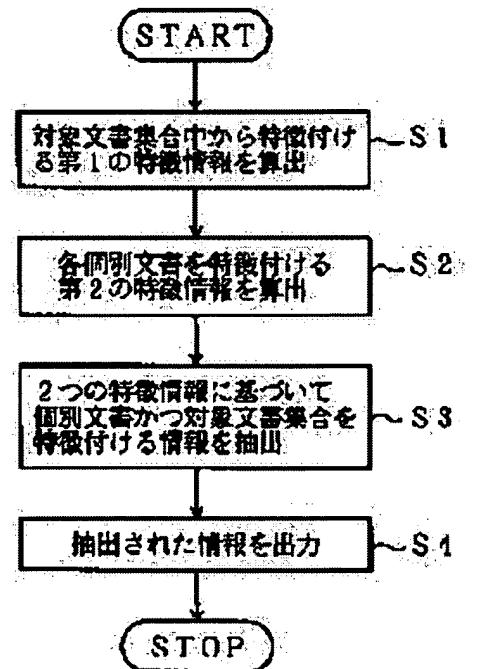
(22)Date of filing : 02.06.1998 (72)Inventor : IZUDERA TOSHINORI

(54) METHOD AND DEVICE FOR EXTRACTING DOCUMENT INFORMATION, AND STORAGE MEDIUM STORED WITH DOCUMENT INFORMATION EXTRACTION PROGRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide the method and device for extracting a document by which even new information can be extracted without preliminary preparing a pattern, heuristics, etc., and to provide a storage medium in which a document information extracting program is stored.

SOLUTION: The 1st characteristic information which characterizes an object document set with respect to a standard document set is calculated from the object document set (S1), the 2nd characteristic information which characterizes each separate document in the object document set with respect to other separate documents is calculated from each separate document in the object document set (S2), the information which characterizes the object document set more and also characterizes each separate document with respect to the other separate documents is extracted from the each separate document (S3) based on the 1st and 2nd characteristic information, and the extracted information is outputted as the information that characterizes each separate document.



BEST AVAILABLE COPY

LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平11-345239

(43)公開日 平成11年(1999)12月14日

(51)Int.Cl.⁶
G 0 6 F 17/30

識別記号

F I
G 0 6 F 15/401
15/40 3 1 0 A
3 7 0 A

(21)出願番号 特願平10-153231

(22)出願日 平成10年(1998)6月2日

(71)出願人 000004226
日本電信電話株式会社
東京都千代田区大手町二丁目3番1号
(72)発明者 岩寺 俊哲
東京都新宿区西新宿三丁目19番2号 日本
電信電話株式会社内
(74)代理人 弁理士 伊東 忠彦

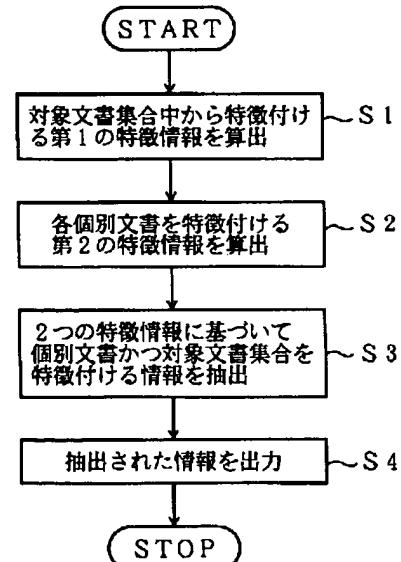
(54)【発明の名称】 文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体

(57)【要約】

【課題】 予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供する。

【解決手段】 本発明は、対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出し、対象文書集合中の各個別文書を他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出し、第1の特徴情報と、前記第2の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、抽出された前記情報を各個別文書を特徴づける情報として出力する。

本発明の原理を説明するための図



【特許請求の範囲】

【請求項1】 文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出方法において、
前記対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出し、
前記対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出し、
前記第1の特徴情報と、前記第2の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、
抽出された前記情報を各個別文書を特徴づける情報として出力することを特徴とする文書情報抽出方法。

【請求項2】 文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出装置であって、
前記対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出する第1の特徴情報算出手段と、
前記対象文書集合中の各個別文書について、他の個別文書を特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出する第2の特徴情報算出手段と、
前記第1の特徴情報算出手段で算出された前記第1の特徴情報と、前記第2の特徴情報算出手段で算出された前記第2の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出手段と、
抽出された前記情報を各個別文書を特徴づける情報として出力する特徴情報出力手段とを有することを特徴とする文書情報抽出装置。

【請求項3】 前記入力記憶手段から標準文書集合を受け取る標準文書集合更新手段と、
前記標準文書集合更新手段に与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析手段と、
前記標準文書集合中の単語と該単語の出現頻度を対応付けて記憶する標準文書集合解析結果記憶手段と、
前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力手段と、
前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析手段と、
前記各対象文書集合中の単語と該単語の出現頻度を各文書を対応付けて記憶する対象文書集合解析結果記憶手段

と、
前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶する個別文書解析結果記憶手段と、

前記対象文書集合全体としての特徴情報を、前記対象文書集合解析結果記憶手段に記憶されている情報を用いて算出する対象文書集合全体特徴算出手段と、

前記対象文書集合全体特徴算出手段によって算出された、前記対象文書集合全体としての特徴情報を記憶する対象文書集合全体特徴記憶手段と、

前記対象文書集合中の各個別文書の特徴情報を、前記個別文書解析結果記憶手段に記憶されている情報を用いて算出する個別文書特徴算出手段と、

前記個別文書特徴算出手段によって算出された、前記対象文書集合中の各個別文書に対応する特徴情報を記憶する個別文書特徴記憶手段と、

前記個別文書解析結果記憶手段または、前記対象文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する目的情報一時記憶手段と、

前記対象文書集合解析結果記憶手段または、前記標準文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する基準情報一時記憶手段と、
前記目的情報一時記憶手段に記憶されているデータと前記基準情報一時記憶手段に記憶されているデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出手段と、

前記対象文書集合全体特徴記憶手段に記憶されているデータと前記個別文書特徴記憶手段に記憶されているデータを用いて、各個別文書の特徴情報を算出する特徴情報算出手段と、

前記特徴情報算出手段において算出された各個別文書の特徴情報を記憶する特徴情報記憶手段と、

前記特徴情報記憶手段に記憶されている各個別文書の特徴情報を用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出手段と、

前記特徴表現抽出手段により前記各個別文書から抽出された特徴表現を記憶する特徴表現記憶手段と、

前記特徴表現記憶手段に記憶されている特徴表現を転送媒体に与える特徴表現出力手段とを有する請求項2記載の文書情報抽出装置。

【請求項4】 前記特徴情報算出手段は、
前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項3記載の文書情報抽出装置。

【請求項5】 前記特徴情報算出手段は、
前記特徴情報を算出する際に χ^2 乗検定を用いる請求項4記載の文書情報抽出装置。

【請求項6】 前記特徴表現抽出手段は、
前記対象文書集合中の各文書中の全単語に、単語の特徴

を数値化した特徴情報スコアを付与する特徴スコア付与手段と、

前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出手段と、

前記平均算出手段により求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定手段とを含む請求項3記載の文書情報抽出装置。

【請求項7】文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出プログラムを格納した記憶媒体であって、

前記対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出する第1の特徴情報算出プロセスと、

前記対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出する第2の特徴情報算出プロセスと、前記第1の特徴情報算出プロセスで算出された前記第1の特徴情報と、前記第2の特徴情報算出プロセスで算出された前記第2の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出プロセスと、

抽出された前記情報を各個別文書を特徴付ける情報として出力する特徴情報出力プロセスとを有することを特徴とする文書情報抽出プログラムを格納した記憶媒体。

【請求項8】前記入力記憶手段から標準文書集合を受け取る標準文書集合更新プロセスと、

前記標準文書集合更新プロセスに与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中の出現頻度を算出する標準文書集合解析プロセスと、

前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、

前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中の出現頻度を算出する対象文書集合解析プロセスと、

前記対象文書集合全体としての特徴情報を、前記各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報を用いて算出する対象文書集合全体特徴算出プロセスと、

前記対象文書集合中の各個別文書の特徴情報を、前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書集合解析

結果記憶手段の情報を用いて算出する個別文書特徴算出プロセスと、

前記個別文書の解析結果または、前記対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、前記対象文書集合解析プロセスまたは、前記標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、

前記対象文書集合全体特徴算出プロセスによって算出された、前記対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、前記個別文書特徴算出プロセスによって算出された、前記対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、

前記特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報を用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、

前記特徴表現抽出プロセスにより前記各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する請求項7記載の文書情報抽出プログラムを格納した記憶媒体。

【請求項9】前記特徴情報算出プロセスは、前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項8記載の文書情報抽出プログラムを格納した記憶媒体。

【請求項10】前記特徴情報算出プロセスは、前記特徴情報を算出する際に χ^2 乗検定を用いる請求項9記載の文書情報抽出プログラムを格納した記憶媒体。

【請求項11】前記特徴表現抽出プロセスは、前記対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、

前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出プロセスと、

前記平均算出プロセスにより求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定プロセスとを含む請求項8記載の文書情報抽出プログラムを格納した記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、文書情報抽出方法

及び装置及び文書情報抽出プログラムを格納した記憶媒体に係り、特に、文書情報処理に用いられる文書情報抽出処理において、特定の話題についての文書集合中の各文書より、当該話題に関連し、かつ、文書集合中の他文書との相違をより明確に示す情報を抽出する文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体に関する。

【0002】

【従来の技術】近年、インターネットが急速に普及している。さらに、データ記憶装置は、大容量化、低価格化している。これに伴って大量で多様な情報が、ネットワークを介して容易に利用可能になっている。また、WWWの普及と共に多くのユーザが相互に情報を生成し、利用している。しかし、情報洪水と言われるように利用できる情報量が飛躍的に増加するに従って、これらの情報の中から有益な情報を見つけ出して取捨選択することが困難になってきている。

【0003】このような大量の情報を全て閲覧し、有益な情報を探索し、選別することは困難である。従って、適切な情報を効率的に利用するためには、大量の情報から特徴的な情報を抽出し、必要十分な情報を選択的に利用可能にする必要がある。現在、情報を選択的に利用する手段として情報検索技術が用いられている。しかし、ネットワークを介して利用できる情報は、その分量が膨大であり、その内容も多岐に渡っている。このため、一度の検索結果として多くの類似の情報を含む文書が選択されてしまう。

【0004】さらに、検索結果から適切な情報を含む文書を選択することを支援するために、多くの場合、各文書を特徴付ける部分情報を文書から抽出し、各検索結果文書に付与し、利用者に提示している。膨大な数の文書の各文書から適量・適質な特徴的な部分情報の、人手による抽出は、困難である。また、人手により作業を行った場合、抽出される情報は、作業者のもつている主観や知識に影響されるため、複数の作業者によって抽出が行われると抽出された情報の品質を均質に保つことができない。そこで、情報を選択的に利用するためには、文書内容から適切な部分情報を自動的に決定し、抽出する情報抽出技術が必要となる。

【0005】従来の情報抽出技術の代表的なものとして次の2つが挙げられる。第1には、パターンマッチングによる情報抽出であり、第2には、ヒューリスティックスを用いた情報抽出技術がある。パターンマッチングによる情報抽出技術は、ある単語列をパターンとして予め保持しておく、パターンマッチング処理によって情報を抽出する技術である。これは、特定の情報が、限られたパターンによって表現されることが多いという考え方に基いている。

【0006】例えば、予め、「<メーカー>は、<製品>の販売を開始した」のようなパターンを用意しておくこ

とにより、文書中からこのパターンにマッチする「〇×コンピュータは新型コンピュータの販売を開始した。」という文を抽出する。また、ヒューリスティックスを用いた情報抽出技術は、文の文書中での位置情報、タイトルや見出しに出現する単語、手がかり語句の有無を組み合わせて文の重要度を判定し、重要と判定された文を抽出する技術である。

【0007】

【発明が解決しようとする課題】しかしながら、上記従来のパターンマッチングによる情報抽出技術は、必要とするパターンを予め用意しておくことが必要である。このため、パターンマッチングしない新たな情報の抽出を行うことができないという問題がある。また、同一の文書からは、必ず同一の部分が抽出される。

【0008】また、上記従来のヒューリスティックスを用いた情報抽出技術は、前述のパターンマッチングによる情報抽出技術と同様に、予めヒューリスティックスを用意しておくことが必要である。このため、新たな情報の抽出には、不向きである。また、同一の文書からは、必ず同一の部分が抽出される。さらに、ある文書タイプで有効なヒューリスティックスが別の文書タイプで有効であるとは限らない。例えば、新聞記事などでは、位置情報が有効である。また、学術論文では、手がかり語句によるヒューリスティックスが有効である。インターネット上には、様々なタイプの文書が混在しており、文書タイプを自動的に判定することが必要となる。しかし、現状では、文書タイプを判別する有効な技術がない、という問題がある。

【0009】また、タイトルや見出しが存在しない文書や、手がかり語句が殆ど出現しない文書も多いため、ヒューリスティックスが有効に働くことが多いという問題がある。本発明は、上記の点に鑑みなされたもので、予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供することを目的とする。

【0010】

【課題を解決するための手段】図1は、本発明の原理を説明するための図である。本発明（請求項1）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出方法において、対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出し（ステップ1）、対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出し（ステップ2）、第1の特徴情報と、第2の特徴情報を基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書か

ら抽出し（ステップ3）、抽出された情報を各個別文書を特徴づける情報として出力する（ステップ4）。

【0011】図2は、本発明の原理構成図である。本発明（請求項2）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出装置であって、対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出する第1の特徴情報算出手段1と、対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出する第2の特徴情報算出手段2と、第1の特徴情報算出手段1で算出された第1の特徴情報と、第2の特徴情報算出手段2で算出された第2の特徴情報に基づいて、対象文書集合をより特徴付ける情報を抽出する個別文書特徴抽出手段3と、抽出された情報を各個別文書を特徴づける情報として出力する特徴情報出力手段4とを有する。

【0012】本発明（請求項3）は、入力記憶手段から標準文書集合を受け取る標準文書集合更新手段と、標準文書集合更新手段に与えられた標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中の出現頻度を算出する標準文書集合解析手段と、標準文書集合中の単語と該単語の出現頻度を対応付けて記憶する標準文書集合解析結果記憶手段と、入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力手段と、対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中の出現頻度を算出する対象文書集合解析手段と、各対象文書集合中の単語と該単語の出現頻度を対応付けて記憶する対象文書集合解析結果記憶手段と、各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶する個別文書解析結果記憶手段と、対象文書集合全体としての特徴情報を、対象文書集合解析結果記憶手段及び標準文書集合解析結果記憶手段に記憶されている情報を用いて算出する対象文書集合全体特徴算出手段と、対象文書集合全体特徴算出手段によって算出された、対象文書集合全体としての特徴情報を記憶する対象文書集合全体特徴記憶手段と、対象文書集合中の各個別文書の特徴情報を、個別文書解析結果記憶手段及び対象文書集合解析結果記憶手段に記憶されている情報を用いて算出する個別文書特徴算出手段と、個別文書特徴算出手段によって算出された、対象文書集合中の各個別文書に対応する特徴情報を記憶する個別文書特徴記憶手段と、個別文書解析結果記憶手段または、対象文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する目的情報一時記憶手段と、対象文書集合解析結果記憶手段または、標準文書集合解析結果記憶手段に記憶されているデータを一時的

に記憶する基準情報一時記憶手段と、目的情報一時記憶手段に記憶されているデータと基準情報一時記憶手段に記憶されているデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出手段と、対象文書集合全体特徴記憶手段に記憶されているデータと個別文書特徴記憶手段に記憶されているデータを用いて、各個別文書の特徴情報を算出する特徴情報算出手段と、特徴情報算出手段において算出された各個別文書の特徴情報を記憶する特徴情報記憶手段と、特徴情報記憶手段に記憶されている各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出手段と、特徴表現抽出手段により各個別文書から抽出された特徴表現を記憶する特徴表現記憶手段と、特徴表現記憶手段に記憶されている特徴表現を転送媒体に与える特徴表現出力手段とを有する。

【0013】本発明（請求項4）は、特徴情報算出手段において、対象文書集合全体特徴と個別文書特徴の特徴スコアを掛けた数値を用いる。本発明（請求項5）は、
20 特徴情報算出手段において、特徴情報を算出する際に χ^2 乗検定を用いる。本発明（請求項6）は、特徴表現抽出手段において、対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与手段と、各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている特徴情報スコアの平均を求める平均算出手段と、平均算出手段により求められた平均の値が最大30 の文書内部分表現を文書の特徴表現として抽出する特徴表現決定手段とを含む。

【0014】本発明（請求項7）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出プログラムを格納した記憶媒体であって、対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出する第1の特徴情報算出プロセスと、対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出する第2の特徴情報算出プロセスと、第1の特徴情報算出プロセスで算出された第1の特徴情報と、第2の特徴情報算出プロセスで算出された第2の特徴情報に基づいて、対象文書集合をより特徴付ける情報を抽出する個別文書特徴抽出プロセスと、抽出された情報を各個別文書を特徴づける情報を出力する特徴情報出力プロセスとを有する。

【0015】本発明（請求項8）は、入力記憶手段から50 標準文書集合を受け取る標準文書集合更新プロセスと、

標準文書集合更新プロセスに与えられた標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中の出現頻度を算出する標準文書集合解析プロセスと、入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の出現頻度を算出する対象文書集合解析プロセスと、対象文書集合全体としての特徴情報報を、各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報を用いて算出する対象文書集合全体特徴算出プロセスと、対象文書集合中の各個別文書の特徴情報を、各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書解析結果記憶手段の情報を用いて算出する個別文書特徴算出プロセスと、個別文書の解析結果または、対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、対象文書集合解析プロセスまたは、標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、対象文書集合全体特徴算出プロセスによって算出された、対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、個別文書特徴算出プロセスによって算出された、対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、特徴表現抽出プロセスにより各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する。

【0016】本発明（請求項9）は、特徴情報算出プロセスにおいて、対象文書集合全体特徴と個別文書特徴の特徴スコアを掛けた数値を用いる。本発明（請求項10）は、特徴情報算出プロセスにおいて、特徴情報を算出する際に χ^2 乗検定を用いる。本発明（請求項11）は、特徴表現抽出プロセスにおいて、対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている特徴情報スコアの平均を求める平均算出プロセスと、平均算出プロセスに

より求められた平均の値が最大の文書内部表現を文書の特徴表現として抽出する特徴表現決定プロセスとを含む。

【0017】上述のように、本発明は、対象文書集合を標準文書集合に対して特徴付ける情報を対象文書集合中から算出し、次に、対象文書集合中の各個別文書について、他の個別文書に対してその文書を特徴付ける情報を個別文書から算出し、これらの処理により得られた特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を各個別文書から抽出する。抽出された情報を各個別文書を特徴付ける情報として出力する。

【0018】これにより、例えば、「桜の花見」について検索を実行した結果、得られる文書集合を「対象文書集合」とし、この文書検索装置が検索対象とする文書集合全体を「標準文書集合」とした場合を想定すると、この場合、対象文書集合を標準文書集合に対して特徴付ける情報は、「桜の花見」に関する情報、例えば、桜の見頃や名所に関する情報になる。

【0019】また、各個別文書を他の個別文書に対して特徴付ける情報は、他の文書と差異がある情報、例えば、「4月の上旬」のような時間的な情報や、「上野公園」のような場所的情報になる。これらの情報を組み合わせて用いることにより、各個別文書からその文書を特徴付ける表現、例えば、「…見頃は、4月上旬…」、「…名所：上野公園…」のような時間的、場所的記述を表している部分を抽出することが可能となる。

【0020】このように、本発明では、大量の文書情報から各文書を特徴付ける情報を適時に適質適量抽出可能となる。また、予め情報抽出に使用する知識等を用意する必要がなく、様々な文書から情報抽出に適用可能である。

【0021】

【発明の実施の形態】図3は、本発明の文書情報抽出装置の構成を示す。同図に示す構成は、例えば、文書検索システムの一部を構成し、検索された文書集合中の各文書から各文書間の相違を明確に示し、各文書を特徴付ける適切な情報を抽出し、提示することにより、使用者が必要とする文書を選択することを支援する装置である。

以下、文書検索システム本体から出力された文書集合中の各文書から情報抽出する場合を想定して説明する。ここで、当該文書情報抽出装置は、入力として文書集合を受信し、予め、提供されている初期情報を用いて入力された文書集合を処理し、各文書毎に情報を抽出し、出力するものである。

【0022】同図に示す構成は、監視制御部10、入力記憶装置20、転送媒体30、各種制御処理を実行する処理部101～110及び各種データを記憶する記憶部201～209から構成される。処理部は、標準文書集合更新部101、標準文書集合解析部102、対象文書

集合入力部103、対象文書集合解析部104、対象文書集合全体特徴算出部105、個別文書特徴算出部106、特徴情報算出部107、特徴表現抽出部108、特徴表現出力部109、目的情報特徴スコア算出部110から構成される。

【0023】記憶部は、標準文書集合解析結果記憶部201、対象文書集合解析結果記憶部202、対象文書集合全体特徴記憶部203、個別文書解析結果記憶部204、個別文書特徴記憶部205、特徴情報記憶部206、特徴表現記憶部207、基準情報一時記憶部208、及び目的情報一時記憶部209から構成される。各処理部101～110を総合的に監視制御する監視制御部10に、処理部101～109と、記憶部201～207が接続される。また、対象文書集合全体特徴算出部105と個別文書特徴算出部106には、目的情報特徴スコア算出部110、基準情報一時記憶部208及び目的情報一時記憶部209が接続される。さらに、基準情報一時記憶部208、目的情報一時記憶部209は、目的特徴スコア算出部110にも接続される。

【0024】ここで、各処理部は、例えば、ディジタル電子計算機で構成され、それぞれCPUと、動作プログラムとそれを実行するためのデータを記録するROMと、ワーキングメモリとして用いられるRAMとを備える。なお、全処理部を1つのディジタル電子計算機で構成してもよい。さらに、各記憶部201～209は、例えば、ハードディスクメモリなどのメモリに記憶される。

【0025】また、入力記憶部20には、本装置に与えられる標準文書集合、対象文書集合が一定の順序で記憶されている。入力記憶部20は、半導体メモリ装置、あるいは、ハードディスクやフロッピーディスクによって実現することができる。転送媒体30には、本装置の処理結果が与えられる通信チャネルまたは記録媒体である。

【0026】以下の説明において、「標準文書集合」とは、特徴情報抽出の対象となり得る文書全体集合、または、文書全体集合を母集合とし、その標本集合となる文書集合を指す。また、「対象文書集合」とは、特徴情報抽出の対象の文書集合を指す。さらに、「個別文書」とは、「対象文書集合」中の各文書を指す。

【0027】「特徴表現」とは、各「個別文書」から抽出される表現であり、「対象文書集合」中の他の文書に対して、当該「個別文書」を特徴付ける表現である。標準文書集合更新部101は、入力記憶装置20から標準文書集合を受け取る。標準文書集合解析部102は、標準文書集合更新部101に与えられた標準文書集合中の各文書を解析し、その文を構成する単語とその単語の標準文書集合中の出現頻度を算出する。

【0028】標準文書集合解析結果記憶部201は、標準文書集合中の単語とその単語の出現頻度を対応付けて

記憶する。対象文書集合入力部103は、入力記憶装置20から複数の文書で構成される対象文書集合を受け取る。対象文書集合解析部104は、対象文書集合中の各文書を解析し、その各個別文書を構成する単語と当該単語の出現頻度を算出する。

【0029】対象文書集合解析結果記憶部202は、対象文書集合解析部104で求められた各対象文書中の単語と単語の出現頻度を各文書と対応付けて記憶する。個別文書解析結果記憶部204は、対象文書集合解析部104で求められた対象文書集合中の各個別文書中の単語とその単語の出現頻度をその単語が出現した文書と対応付けて記憶する。

【0030】対象文書集合全体特徴算出部105は、対象文書集合全体としての特徴情報を、上記対象文書集合解析結果記憶部202に記憶されている情報及び上記標準文書集合解析結果記憶部201に記憶されている情報を用いて算出する。対象文書集合全体特徴記憶部203は、対象文書集合全体特徴算出部105によって算出された、対象文書集合全体としての特徴情報を記憶する。

【0031】個別文書特徴算出部106は、対象文書集合中の各個別文書の特徴情報を個別文書解析結果記憶部204に記憶されている情報及び対象文書集合解析結果記憶部202に記憶されている情報を用いて算出する。個別文書特徴記憶部205は、個別文書特徴算出部106によって算出された対象文書集合中の各個別文書の特徴情報を記憶する。

【0032】目的情報一時記憶部209は、個別文書解析結果または、対象文書集合解析結果を一時的に記憶する。基準情報一時記憶部208は、対象文書集合解析結果、または、標準文書集合解析結果を一時的に記憶する。目的情報特徴スコア算出部110は、目的情報一時記憶部209に記憶されているデータと上記基準情報一時記憶部208に記憶されているデータを比較し、上記の目的情報一時記憶部209に記憶されているデータの特徴スコアを算出する。

【0033】特徴情報算出部107は、対象文書集合全体特徴記憶部203に記憶されているデータと個別文書特徴記憶部205に記憶されているデータを用いて各個別文書の特徴情報を算出する。特徴情報記憶部206は、特徴情報算出部107において算出された各個別文書の特徴情報を記憶する。

【0034】特徴表現抽出部108は、特徴情報記憶部206に記憶されている各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する。特徴表現記憶部207は、特徴表現抽出部108により各個別文書から抽出された特徴表現を記憶する。特徴表現出力部109は、特徴表現記憶部207に記憶されている特徴表現を転送媒体30に与える。

【0035】
50 【実施例】以下、図面と共に本発明の実施例を説明す

る。まず、監視制御部10及び対象文書集合全体特徴算出部105、個別文書特徴算出部106に接続される記憶部201～209について説明する。標準文書集合解析結果記憶部201は、標準文書集合解析部102の処理結果である、標準文書集合の解析結果を記憶・保持する。標準文書集合の解析結果とは、標準文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現及び各単語出現頻度を対応付けたものである。解析結果は、単語表現、出現頻度の2つのカラムからなるテーブルとして表現・記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【0036】対象文書集合解析結果記憶部202は、対象文書集合解析部104の処理結果の1つとして得られる対象文書集合の解析結果を記憶・保持する。対象文書集合とは、本装置が接続される情報検索装置において検索作業の実行を結果として得られる文書集合である。対象文書集合の解析結果とは、対象文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現と対象文書集合中の各単語の出現頻度を対応付けたものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の2つのカラムからなるテーブルとして、表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【0037】個別文書解析結果記憶部204は、対象文書集合解析部104の処理結果の1つとして得られる個別文書の解析結果を記憶・保持する。ここで、個別文書とは、本装置が接続される情報検索装置において、検索作業の実行の結果として得られる文書集合、即ち、前述した対象文書集合に含まれる各文書である。個別文書の解析結果とは、対象文書集合中の各文書毎に記述されている文章を形態素解析し、各単語の表現、及びその文書中の各単語の出現頻度を対応付け、文書毎に記録したものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の2つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。また、対象文書集合中の各文書毎に1個のテーブルが構成される。

【0038】対象文書集合全体特徴記憶部203は、対象文書集合全体特徴算出部105の処理結果として得られる、標準文書集合に対する対象文書集合全体として特徴を点数化した情報を記憶・保持する。ここで、対象文書集合全体特徴とは、標準文書集合中の単語の出現頻度分布と対象文書集合中の単語の出現頻度分布を比較し、

その分布の相違の大小を各単語毎に数値化したものであり、標準文書集合中の出現頻度分布と対象文書集合中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。対象文書集合中で特徴的な単語、即ち、標準文書集合中の出現頻度分布と対象文書集合中の出現分布の相違が大きい単語ほど大きな数値をとる。対象文書集合全体特徴は、各単語表現とその単語の出現分布の特徴を数値化し、表現した特徴スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。対象文書集合全体に対して1つのテーブルが対応する。

【0039】個別文書特徴記憶部205は、個別文書特徴算出部106の処理結果として得られる。対象文書集合全体に対する対象文書集合中の各個別文書の特徴を点数化し、記憶・保持する。ここで、個別文書特徴とは、対象文書集合中の単語の出現頻度分布と対象文書集合中の各個別文書中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、対象文書集合中の出現分布と個別文書中の出現分布の相違が大きい単語ほど、大きな数値をとる。即ち、個別文書（対象文書集合中の各文書）に特徴的な単語ほどより大きな数値を持つ。個別文書特徴は、各単語表現とその単語の出現分布の特徴を数値化し表現した特徴スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。また、個別文書毎に1つのテーブルが対応する。

【0040】特徴情報記憶部206は、特徴情報算出部107の処理結果として得られる、各文書の特徴情報をその情報の算出元である文書と対応付けて記憶・保持する。ここで、特徴情報とは、各単語毎に対象文書集合中の特徴スコアと個別文書中の特徴スコアをかけた数値であり、対象文書集合中に特徴的な単語であり、かつ個別文書中でも特徴的な単語ほど大きな数値を持つ。特徴情報は、各単語表現とその単語の特徴を数値化した特徴スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴スコアの対応を表す。また、個別文書毎に1つのテーブルが対応する。

【0041】特徴表現記憶部207は、特徴表現抽出部108の処理結果として得られる、対象文書集合中の各文書毎の特徴表現をその情報の抽出元である文書と対応付けて記憶・保持する。各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた部分構造を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列を構成する単語全体の特徴スコアの平均が最大の部分である。

【0042】基準情報一時記憶部208は、目的情報特徴スコア算出部110における特徴スコアの算出に用いる基準情報を一時的に記憶・保持する。基準情報は、単語表現、出現頻度の2つのカラムからなるテーブルとし

て表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部208には、一度に上記の形式のテーブルが1個、記憶・保持される。

【0043】目的情報一時記憶部209は、目的情報特徴スコア算出部110における特徴スコアの算出において、特徴スコアの算出の対象となる目的情報を一時的に記憶・保持する。目的情報は、単語表現、出現頻度の2つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部209には、一度に上記形式のテーブルが1個、記憶・保持される。

【0044】次に、図3に示す各処理部について説明する。監視制御部10は、処理部101～109を制御し、データフローを統制するモジュールである。図4は、本発明の一実施例の文書情報抽出処理のフローチャートである。以下、同図のフローチャートに沿って、各処理部の動作を説明する。

【0045】ステップ101）監視制御部10において、標準文書集合が更新されているか否かが判断される。更新された場合には、ステップ102に移行し、更新されていない場合は、ステップ105に移行する。

ステップ102）監視制御部10は、更新された標準文書集合を入力記憶装置20から標準文書集合更新部101へ転送する。

【0046】ステップ103）この時点で、標準文書集合更新部101は、転送された標準文書集合に対して、標準文書集合更新処理を実行し、処理結果である標準文書集合更新結果を監視制御部10に出力する。監視制御部10は、標準文書集合更新部101から出力されたすべての標準文書集合更新結果を標準文書集合解析部102へ転送する。

【0047】ステップ104）標準文書集合解析部102は、標準文書集合解析処理を実行し、処理結果を監視制御部10へ出力する。これにより、監視制御部10は、標準文書集合解析部102から出力されたすべての標準文書集合解析結果を標準文書集合解析結果記憶部201に転送し、その内容を更新し、新たに転送された値を記憶・保持する。

【0048】ステップ105）監視制御部10によって対象文書集合が入力されたか否かが判断される。入力された場合は、ステップ106へ移行し、入力されていない場合には、ステップ105の処理を繰り返す。

ステップ106）監視制御部10は、入力された対象文書集合を入力記憶装置20から対象文書集合入力部103へ転送する。対象文書集合入力部103は、入力された対象文書集合に対して対象文書集合入力処理を実行

し、処理結果を監視制御部10に出力する。監視制御部10は、対象文書集合入力部103から出力される対象文書集合入力処理結果を対象文書集合解析部104へ転送する。

【0049】ステップ107）対象文書集合解析部104は、対象文書集合解析処理を実行し、解析結果を監視制御部10へ出力する。

ステップ108）監視制御部10は、対象文書集合解析部104から出力された対象文書集合解析結果を対象文書集合解析結果記憶部202に転送すると共に、個別文書解析結果を個別文書解析結果記憶部203に転送

10し、各記憶部の内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、標準文書集合解析結果記憶部201に記憶されている標準文書集合解析結果を対象文書集合全体特徴算出部105へ転送すると共に、対象文書集合解析結果記憶部202に記憶されている対象文書集合解析結果を対象文書集合全体特徴算出部105へ転送する。

【0050】ステップ109）対象文書集合全体特徴

20算出部105は、監視制御部10から転送された標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴算出処理を実行し、処理結果を監視制御部10に出力する。監視制御部10は、対象文書集合全体特徴算出部105から出力された対象文書集合全体特徴を対象文書集合全体特徴記憶部203に転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、対象文書集合解析結果記憶部202に記憶されている対象文書集合解析結果を個別文書特徴算出部106へ転送すると共に、個別文書解析結果記憶部204に記憶されている個別文書集合解析結果を個別文書特徴算出部106へ転送する。

【0051】ステップ110）個別文書特徴算出部106は、転送されてきた対象文書集合解析結果と個別文書解析結果に基づいて個別文書特徴算出処理を実行す

る。処理結果である、個別文書特徴情報は、監視制御部10に出力する。監視制御部10は、個別文書特徴算出部106から出力された個別文書特徴情報を個別文書特徴記憶部205に転送し、その内容を更新し、新たに転送された値を記憶・保持する。監視制御部10は、対象文書集合全体特徴記憶部203に記憶されている対象文書集合全体特徴を特徴情報算出部107へ転送する。それと共に、個別文書特徴記憶部205に記憶されている個別文書特徴を特徴情報算出部107へ転送する。

【0052】ステップ111）特徴情報算出部107は、対象文書集合全体特徴と個別文書特徴に基づいて特徴情報算出処理を実行し、処理結果である特徴情報を監視制御部10へ出力する。監視制御部10は、特徴情報算出部107から出力された特徴情報を特徴情報記憶部206へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、特徴

情報記憶部206に記憶されている特徴情報を特徴表現抽出部108へ転送する。

【0053】ステップ112) 特徴表現抽出部108は、転送されてきた特徴情報に基づいて、特徴表現抽出処理を実行し、処理結果である特徴表現を、監視制御部10へ出力する。監視制御部10は、特徴表現抽出部108から出力された特徴表現を特徴表現記憶部207へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、特徴表現記憶部207に記憶されている特徴表現を特徴表現出力部109へ転送する。

【0054】ステップ113) 特徴表現出力部109は、転送されてきた特徴表現について特徴表現出力処理を実行する。処理結果は、監視制御部10へ出力され、監視制御部10において、特徴情報出力部109から出力された特徴情報出力処理結果を転送媒体30へ出力する。

ステップ114) すべての処理が終了か否かを判定し、すべての処理が終了している場合には、当該監視制御処理を終了する。また、終了していない場合には、ステップ105に移行し、上述の処理を繰り返す。

【0055】以下に、各部の詳細な処理について説明する。標準文書集合更新部101では、監視制御部10から転送された標準文書集合に対して標準文書集合更新処理が実行される。この処理は、以降の処理の前処理であり、入力された標準文書集合中の各文書から本装置による処理に必要な部分を除去する。また、以降の処理で対応している文字コードへ変換される。処理結果は、監視制御部10へ出力される。

【0056】標準文書集合解析部102では、監視制御部10から転送される標準文書集合更新処理結果に対して標準文書集合解析処理が実行される。この処理は、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び転送されてきた標準文書集合中の各単語の出現頻度を対応付けて記録するものである。解析結果は、単語表現、出現頻度の2つのカラムからなるテーブルとして監視制御部10へ出力される。このテーブルにおいて、各行には、各単語表現、及びその単語の出現頻度が記述される。また、このテーブルは、各単語表現をキーとして対応する全ての行を検索できる構造をとる。

【0057】対象文書集合入力部103では、監視制御部10から転送されてくる対象文書集合に対して対象文書集合入力処理が実行される。この処理は、以降の処理の前処理であり、転送された対象文書集合中の各文書から本装置による処理に必要な部分を除去する。また、以降の処理で対応している文字コードへ変換される。処理結果は監視制御部10へ出力される。

【0058】対象文書集合解析部104では、監視制御部10から転送されてくる対象文書集合入力処理結果に

対して対象文書集合解析処理が実行され、処理結果として対象文書集合解析結果、個別文書解析結果が監視制御部10へ出力される。対象文書集合解析処理は、まず、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び、その文書中の各単語出現頻度を対応付け、文書毎のテーブルに記録する。このテーブルは、対象文書集合中のすべての文書に対して、文書毎に1個作られる。これらのテーブルが、個別文書解析結果として監視制御部10へ出力される。次に、これらのすべてのテーブルに対して同一単語表現の出現頻度が合算され、対象文書集合全体に対して、1個のテーブルが作られる。このテーブルが対象文書集合解析結果として監視制御部10へ出力される。上記のすべてのテーブルは、単語表現を記録するカラムと、その単語の出現頻度を記録するカラムの2つのカラムから構成される。また、このテーブルは、各単語表現をキーとして対応するすべての行を検索できる構造をとる。

【0059】対象文書集合全体特徴算出部105では、監視制御部10から転送されてくる、標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴を算出する。対象文書集合全体特徴の算出手順は、次のようにになる。

① 標準文書集合解析結果を基準情報一時記憶部208に転送する。

【0060】② 対象文書集合解析結果を目的情報一時記憶部209に転送する。

③ 目的情報特徴スコア算出部110を起動し、算出結果が返されるのを待機する。

④ 目的情報特徴スコア算出部110から算出結果が返されると、それを対象文書集合全体特徴として監視制御部10へ出力する。

【0061】個別文書特徴算出部106では、監視制御部10から転送されてくる、対象文書集合解析結果と個別文書解析結果に基づいて、対象文書集合中に含まれている各文書毎に個別文書特徴を算出する。個別文書特徴の算出手順は、以下になる。

① 対象文書集合解析結果を基準情報一時記憶部208に転送する。

【0062】② 個別文書解析結果を目的情報一時記憶部209に転送する。

③ 目的情報特徴スコア算出部110を起動し、算出結果が返されるのを待機する。

④ 目的情報特徴スコア算出部110から算出結果が返されると、それを個別文書特徴として監視制御部10へ出力する。

【0063】特徴情報算出部107では、監視制御部10から転送されてくる、対象文書集合特徴と個別文書特徴に基づいて、対象文書集合中の各文書毎に、特徴情報を算出する。算出結果は、監視制御部10へ出力される。特徴情報は、各単語表現とその単語の特徴を数値化

した特徴情報スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴情報スコアの対応を表す。

【0064】また、個別文書毎に1つのテーブルが構成される。特徴情報算出の手順は、次のようになる。

① 各単語毎に対象文書集合特徴中での特徴スコアと個別文書特徴中の特徴スコアを読み出す。

② これらのスコアの積を求める。この結果が各単語の特徴情報スコアとなる。

【0065】対象文書集合中に特徴的単語であり、かつ、個別文書中でも特徴的な単語ほど、この特徴情報スコアは大きな数値を持つ。特徴表現抽出部108では、監視制御部10から転送されてくる特徴情報を用いて、対象文書集合中の各文書から各文書毎にその文書に特徴的な表現（特徴表現）を抽出し、その表現が抽出された文書と対応付け、監視制御部10へ出力する。

【0066】各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造（例えば、段落）を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列に含まれる各単語の特徴情報スコアの平均が最大の部分である。

【0067】特徴表現として、各文書から一文を抽出する場合、特徴表現抽出の手順は以下のようになる。

① まず、文書中の全単語に特徴情報スコアを付与する。

② 次に、文書中の各文毎に、その文を構成する単語に付与されている特徴情報スコアの平均を求める。

【0068】③ 特徴情報スコアの平均が最大の文をその文書の特徴表現として抽出する。特徴表現出力部109では、監視制御部10から転送されてくる、各個別文書に対応する特徴表現を監視制御部10を通して転送媒体30に出力する。目的情報特徴スコア算出部110では、基準情報一時記憶部208と目的情報一時記憶部209に各々記憶・保存されている基準情報と目的情報に基づいて、特徴スコアを算出し、出力する。

【0069】特徴スコアとは、基準情報中の単語の出現頻度分布と目的情報中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、基準情報中の出現頻度分布と目的情報中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。即ち、目的情報に特徴的な単語ほど、より大きな数値を持つ。

【0070】特徴スコアは、各単語毎にその単語の出現頻度分布に対して、 χ^2 乗検定の考え方を用いて算出する。 χ^2 乗検定は、「いくつかの群で、ある変数の分布に差があるかどうか」を検定することができる。本発明では、この変数を文書中の単語とする。例えば、標準文書集合に対する対象文書集合の特徴スコアを算出する場合、対象文書集合と標準文書集合中の全単語の出現総数

と対象文書集合中の全単語の出現総数と標準文書集合中の全単語の出現総数から計算される各単語の各文書集合中での出現頻度の期待値の分布と、実際に観測される各単語の各文書集合中での出現頻度の分布から χ^2 乗値を算出する。この値が大きくなるほど分布に差があることになり、そのような単語ほど偏って出現していることになる。本発明では、この値を用いて各単語の特徴スコアを算出する。

【0071】以下に標準文書集合と対象文書集合が与えられている場合の具体的な例を用いて説明する。図5は、本発明の一実施例の標準文書集合の一部の例を示し、図6は、本発明の一実施例の対象文書集合の一部の例を示す。以下に各処理部における詳細な処理動作を前述のフローチャートに基づいて説明する。

【0072】まず、標準文書集合が更新されているか否かが判定される（ステップ101）。この例では更新されていると判明したものとする。図5に示す標準文書集合が、標準文書集合更新部101に転送される（ステップ102）。なお、標準文書集合は、文書集合であり、

20 その文書数は十分に大きいことが望ましい。標準文書集合更新部101では、標準文書集合中の各文書から以後の処理に不要である部分が除去される（ステップ103）。例えば、HTML形式の文書の場合は、HTMLタグが除去される。また、ワープロ文書の場合は、文字飾り等が除去される。さらに、文書を構成している文字のコードがまちまちである場合は、1つのコードに統一される。この結果は、監視制御部10に出力される。監視制御部10は、これを標準文書集合解析部102に転送する。

30 【0073】標準文書集合解析部102は、転送されてきた標準文書集合を解析する（ステップ104）。即ち、各文書毎にその文書が記述している文章を形態素解析し、標準文書集合中の単語表現とその単語の出現頻度を求める。図7は、本発明の一実施例の標準文書集合解析結果の例を示す。同図に示す標準文書集合解析結果は、監視制御部10に出力される。監視制御部10は、標準文書集合解析結果を標準文書集合解析結果記憶部201に転送する。標準文書集合解析結果記憶部201は、転送されてきた標準文書集合解析結果を記憶・保持する。

40 【0074】以上により、標準文書集合に関する処理が完了する。次に、対象文書集合が入力されているか否かが判定される（ステップ105）。入力されている場合は、以下のように処理が進行する。図6に示す対象文書集合が、対象文書集合入力部103に入力される（ステップ106）。対象文書集合入力部103は、入力された文書集合から以降の処理に不要の部分を除去する。また、以降の処理に対応する文字コードへ変換する。処理結果は、監視制御部10に出力される。監視制御部10は、対象文書集合入力部103から出力された対象文書

集合を対象文書集合解析部104に転送する。

【0075】対象文書集合解析部104は、転送された対象文書集合を解析する(ステップ107)。解析結果は対象文書集合解析結果と個別文書解析結果である。対象文書集合解析結果は、対象文書集合を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図8に示す。また、個別文書解析結果は、対象文書集合を構成する各文書毎のその文書を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図9に示す。これらの結果は監視制御部10に出力される。

【0076】監視制御部10は、対象文書集合解析結果を対象文書集合解析結果記憶部202に、個別文書解析結果を個別文書解析結果記憶部204にそれぞれ転送する。対象文書集合解析結果記憶部202は、転送されてきた対象文書集合解析結果を記憶・保持する。同様に、個別文書解析結果記憶部204は、転送されてきた個別文書解析結果を記憶・保持する(ステップ108)。

【0077】次に、対象文書集合全体特徴を計算する(ステップ109)。まず、監視制御部10は、標準文書集合解析結果記憶部201に記憶・保持されている標準文書集合解析結果と、対象文書集合解析結果記憶部202に記憶・保持されている対象文書集合解析結果を対象文書集合全体特徴算出部105に転送する。対象文書集合全体特徴算出部105は、転送されてきた標準文書集合解析結果を基準情報一時記憶部208に、同様に、転送されてきた対象文書集合解析結果を目的情報一時記憶部209に転送する。

【0078】次に、特徴スコア算出部110を起動する。特徴スコア算出部110では、基準情報一時記憶部208と目的情報一時記憶部209を参照して特徴スコアを算出する。算出結果は、対象文書集合全体特徴算出部105に出力させる。図10は、特徴スコアの算出結果を示す。対象文書集合全体特徴算出部105では、特徴スコア算出部110の出力結果を対象文書集合全体特徴として図10に示すような結果を監視制御部10に出力する。監視制御部10は、対象文書集合全体特徴記憶部204に転送する。対象文書集合全体特徴記憶部204は、転送されてきた対象文書全体特徴を記憶・保持する。

【0079】次に、対象文書集合中の各文書毎に、個別文書特徴を計算する(ステップ110)。まず、監視制御部10は、対象文書集合解析結果記憶部202に記憶・保持されている対象文書集合解析結果を個別文書特徴算出部106に転送する。また、監視制御部10は、個別文書解析結果記憶部204に記憶・保持している個別文書解析結果を、一文書分毎に個別文書特徴算出部106に転送する。

【0080】次に、特徴スコア算出部110を起動する。特徴スコア算出部110では、基準情報一時記憶部

208と目的情報一時記憶部209を参照して特徴スコアを算出する。算出結果は、個別文書特徴算出部106に出力される。その結果の一部を図11に示す。個別文書特徴算出部106では、目的特徴スコア算出部110の出力結果を個別文書特徴として監視制御部10に出力する。その結果一部を図11に示す。監視制御部10は、個別文書特徴算出部106で得られた個別文書特徴を個別文書特徴記憶部205に転送する。個別文書特徴記憶部205は、転送されてきた個別文書特徴を記憶・保持する。

【0081】次に、対象文書集合中の各文書毎に特徴情報算出部111)。まず、監視制御部10は、対象文書集合全体特徴記憶部203に記憶・保持されている対象文書集合全体特徴を特徴情報算出部107に転送する。また、同様に、個別文書特徴記憶部205に記憶・保持されている個別文書特徴を各文書毎に特徴情報算出部107に転送する。

【0082】特徴情報算出部107は、監視制御部10から転送されてきた対象文書全体特徴と個別文書特徴を用いて特徴情報を算出し、算出結果を監視制御部10へ出力する。算出結果の一部を図12に示す。ここで、特徴情報とは、各単語毎に対象文書集合中の特徴スコアと個別文書中の特徴スコアを掛けた数値であり、対象文書集合中に特徴的単語であり、かつ、個別文書でも特徴的な単語程大きな数値を持つ。

【0083】監視制御部10は、特徴情報算出部107の出力結果を各文書毎に特徴情報記憶部206に転送する。特徴情報記憶部206は、転送されてきた特徴情報を文書毎に記憶・保持する。次に、対象文書集合中の各文書から特徴表現を抽出する(ステップ112)。まず、監視制御部10は、特徴情報記憶部206に記憶・保持されている特徴情報を特徴表現抽出部108に転送する。

【0084】特徴表現抽出部108では、監視制御部10から転送されてきた特徴情報を用いて各文書から特徴表現を抽出し、その結果を監視制御部10へ出力する。抽出結果を図13に示す。監視制御部10は、特徴表現抽出部108の出力結果を各文書毎に特徴表現記憶部207に転送する。

【0085】特徴表現記憶部207は、監視制御部10から転送されてきた特徴表現を文書毎に記憶・保持すると共に、監視制御部10に出力する。監視制御部10は、特徴表現を特徴表現出力部109に転送する。監視制御部10は、対象文書集合中のすべての文書において、特徴表現抽出を終了後、特徴表現出力部109は、特徴表現記憶部207に記憶・保持されている特徴表現をそれが抽出された文書と対応付けし、転送媒体30に出力する(ステップ113)。

【0086】以上の実施例において、種々の定義値を用いているが、これらの値は設計値であり、下記のように

必要に応じて変更してもよい。

・特徴情報の算出に χ^2 乗検定の考え方を用いているが、他の手法で算出してもよい。
・特徴量スコアの算出単位として単語を用いたが、この単位は文字や一定長の文字列でもよい。

【0087】また、上記の実施例では、図3に示す構成に基づいて説明しているが、この例に限定されることなく、専用のハードウェア回路によって実現することも可能であり、さらに、プログラムされたコンピュータによって実現することも可能である。つまり、監視制御部100、標準文書集合更新部101、標準文書集合解析部102、対象文書集合入力部103、対象文書集合解析部104、対象文書集合全体特徴算出部105、個別文書特徴算出部106、特徴情報算出部107、特徴表現抽出部108、特徴表現出力部109、目的情報特徴スコア算出部110をプログラムとして構築し、文書情報抽出装置として利用されるコンピュータに接続されるディスク装置や、フロッピーディスク、CD-ROM等の可搬記憶媒体に格納しておき、本発明を実施する際に、インストールすることにより、容易に本発明を実現することが可能である。

【0088】なお、本発明は、上記の実施例に限定されることなく、特許請求の範囲内で種々変更・応用が可能である。

【0089】

【発明の効果】上述のように、本発明によれば、予め情報抽出知識やパターンなどを用意することなく、文書情報を抽出することが可能となる。これにより、使用開始時に想定したものと対象とする文書内容に差異が生じた場合や、新たな情報を含んでいる場合においても、適切に文書情報を抽出することが可能である。

【0090】また、文書集合中の各文書を比較するのに適した各文書を特徴付ける文書情報が抽出できるので、文書検索システムの出力編集装置に適用することにより、効率的に検索結果の文書集合から文書を選択、閲覧することができる。

【図面の簡単な説明】

【図1】本発明の原理を説明するための図である。

【図2】本発明の原理構成図である。

【図3】本発明の文書情報抽出装置の構成図である。

【図4】本発明の一実施例の文書情報抽出処理のフローチャートである。

*

* 【図5】本発明の一実施例の標準文書集合の例である。
【図6】本発明の一実施例の対象文書集合の例である。
【図7】本発明の一実施例の標準文書集合解析結果の例である。
【図8】本発明の一実施例の対象文書集合解析結果の例である。
【図9】本発明の一実施例の個別文書解析結果の例である。
【図10】本発明の一実施例の対象文書集合全体特徴の例である。
【図11】本発明の一実施例の個別文書特徴の例である。
【図12】本発明の一実施例の特徴情報の例である。
【図13】本発明の一実施例の特徴表現の例である。
【符号の説明】

- | | |
|-----|---------------|
| 1 | 第1の特徴情報算出手段 |
| 2 | 第2の特徴情報算出手段 |
| 3 | 個別文書特徴抽出手段 |
| 4 | 特徴情報出力手段 |
| 10 | 監視制御部 |
| 20 | 入力記憶装置 |
| 30 | 転送媒体 |
| 101 | 標準文書集合更新部 |
| 102 | 標準文書集合解析部 |
| 103 | 対象文書集合入力部 |
| 104 | 対象文書集合解析部 |
| 105 | 対象文書集合全体特徴算出部 |
| 106 | 個別文書特徴算出部 |
| 107 | 特徴情報算出部 |
| 108 | 特徴表現抽出部 |
| 109 | 特徴表現出力部 |
| 110 | 目的情報スコア算出部 |
| 201 | 標準文書集合解析結果記憶部 |
| 202 | 対象文書集合解析結果記憶部 |
| 203 | 対象文書集合全体特徴記憶部 |
| 204 | 個別文書解析結果記憶部 |
| 205 | 個別文書特徴記憶部 |
| 206 | 特徴情報記憶部 |
| 207 | 特徴表現記憶部 |
| 208 | 基準情報一時記憶部 |
| 209 | 目的情報一時記憶部 |

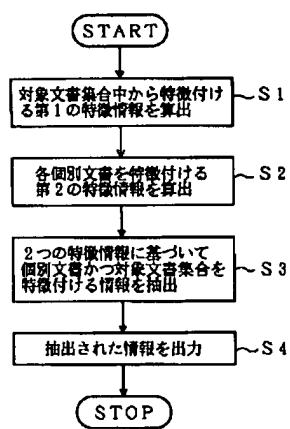
【図13】

本発明の一実施例の特徴表現の例

〔四一〕

〔图2〕

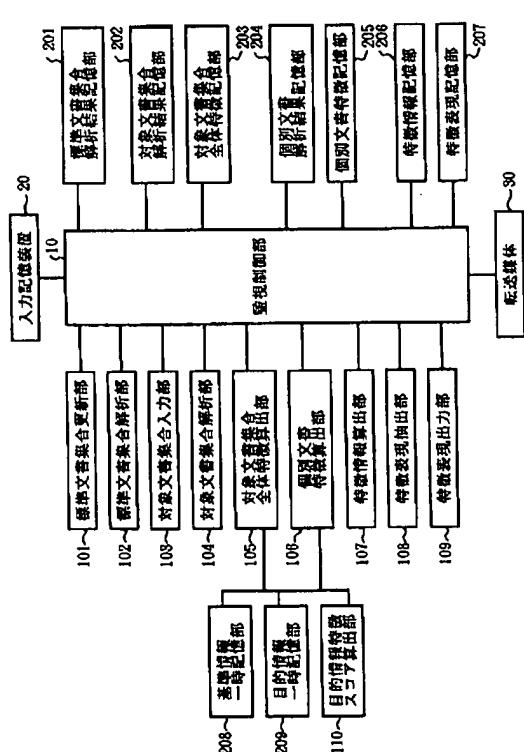
本発明の原理を説明するための図



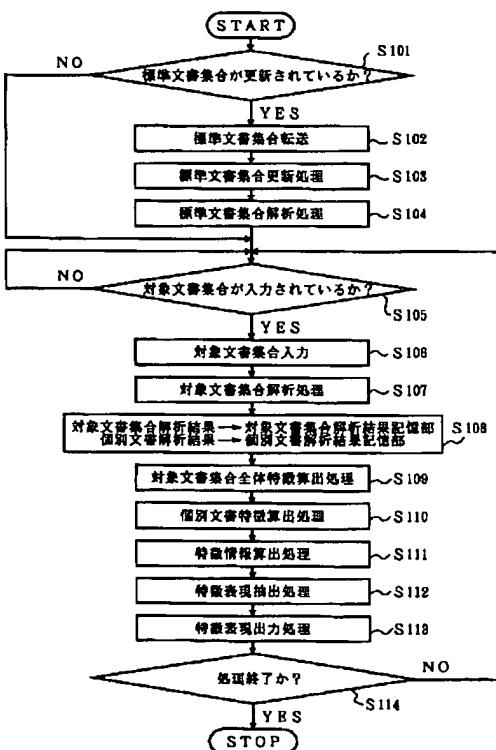
〔图3〕

[図4]

本発明の文書情報抽出装置の構成図



本発明の一実施例の文書情報抽出処理のフローチャート



【図5】

本発明の一実施例の標準文書集合の例

真はどうしても食欲が減退しがちだけど、これは植物。暑い季節を乗り切るためにも一定のスタイルはもちらん重要な要素だから、あるいは野郎のお店に出来てもらいたい。東京の古本ネット販売店は、その前に月を楽しむせで情報を収集して、販売するだけは避けたい。そして、なによりこの季節は、おもとYahoo! JAPANでは引き続きY-hoo's Picks in the Weekで取り上げるにふさわしいお寄せをsugest-picks@yahoo.co.jpで募集中。お待ちしています。

おいしいメニューのちゃんとちゃんとお味が提供すること、では家庭で簡単に設立できるような簡単なメニューを一緒に、自分で作れるどといった目的別に分類して案内している。家庭で作れるシエフの味。では、なんと田舎先生のサケ、チラウマ、朝からえびのチリソースのレシピなど、自分で挑戦されるといいから。おでかけメニューが佐々木さんだけじゃなく、朝はん、朝ごはんを相手に作ってもらいたいな。

【図6】

本発明の一実施例の対象文書集合の例

千鳥ヶ淵周辺
ちどりがぶらしゅうへん
東京都千代田区

千鳥ヶ淵周辺内および周辺のさくらは、田安門跡のものと併せて昭和30年(1955)開設に植えられたもの。後に区内のさくらは、当時街路樹として植えられたといわれる。昭和51年(1976)にこの植えは整備完成した。緑道内の千鳥ヶ淵公園のさくらは首都高速道路建設で一時撤去されたが、若木を再植栽した。

*所在地 東京都千代田区九段南、三番町、九段北、一番町
*交差 地下鉄東西線・新宿線九段下駅または半蔵門線半蔵門駅下車徒歩5分

*見頃 4月1日～14日

*種類状況 ソメイヨシノ、ヤマザクラ、シダレザクラ、サツザクラなど
798本

*開花期 4月1日～16日

*周辺の施設 皇居東御苑、靖国神社、北の丸公園、国立近代美術館、科学技術館

*特徴品 特になし

*問い合わせ先 千代田区企画部広報課 電話03-5284-0151

[Image] 前ページへ 次ページへ [Image]

アブルはこれから12ヶ月かけて総額2千万ドルの予算を確保し、様々なマーケティングやチャネル・マーケティング・プログラムを運営企画におけるマーケティングの有効性をおささげに明確にする。次に、マーケティングの有効性がお水に溶け、アブルはラベロングアーバンマーケティングの実行のために自部門内での協調やはばたきのいいます。よう最近、ハイディ・ロイゼンは、この2つについてのうな見解を述べました。彼女の見解が認知されるよう、これまで月別に書類を提出するが、それが月別に書類を提出する。2を複数の活動に割り当てる。とくにこの計画の目的は、特に当てられるおよその半分を米国以外のアブルの市場に割り当てる。現在、日本の販売網(流通網)と新たな計画ネットワークにあります。この資金は、流通網に投げ込まれるような資金を確保してあります。

アブルは、マーケティングの有効性を確保するためのツールなどに配ります。Mac OS 対応の様々な既存のソリューションといふうはソリューションの幅や多様性をアピールするため使用されます。言い換えれば、個々の製品のマーケティング活動を支援するような特徴。そのため資金を確保するわけではありません。

西公園
にしこうえん
福岡県／福岡市

黒田長政福岡移封と同時に、東田梅原を守護神として崇って以来、蓮山として知られる福岡の城である。明治14年(1881)、公園地となりた折に霧津山公園と呼ばれ。地元有志で組織した元町山武屋会がその管理をまかされ、明治33年(1900)、元町山公園と改められ、5千本、もじなどを見樹木になつた。

*所在地 地下鉄、福岡大橋公園駅下車徒歩10分
*交差 地下鉄、福岡大橋公園駅下車徒歩10分/福岡市高速自動車道西福岡ランプから0.5キロ

*見頃 4月上旬

*種類状況 ソメイヨシノ、サツザクラ、ヤマザクラ、ヒガンザクラなど3190本。(宝くじ桜植栽地)

*周辺の施設 特になし

*開花期 4月上旬

*問い合わせ先 大濠公園、福岡城跡、南公園、市立歴史博物館
福岡市西区
*登録登録 諸多人形、燈籠、燈籠

*問い合わせ先 福岡県福岡公園事務所 電話092-741-2004

次ページへ [Image]

【図7】

本発明の一実施例の標準文書集合解析結果の例

単語	頻度	単語	頻度
ページ	14	圓	2
年	11	用意	2
の	11	評判	2
月	10	活動	3
日	10	販売	2
号	9	味	3
製品	7	分	3
皆さん	7	意見	3
こと	7	目的	3
情報	6	自分	3
メニュー	6	人	3
マーケティング	6	話	3
資金	6	スタッフ	3
英語	5	市	3
歌舞伎	5	広島	5
ジャズ	5	派遣	3
ため	5	アップル	3
計画	5	中心	2
店	5	標準	2
インターネット	4	必なた	2
支援	4	式典	2
哈	4	日本	2
歴史	4	大会	2
紹介	4	通訳	2
ホーム	4	対応	2
部門	4	効果	2
プログラム	4	個々	2
序説	4	生態	2
よう	4	レコーディング	2
米	4	お支障	2

【図8】

本発明の一実施例の対象文書集合解析結果の例

単語	頻度	単語	頻度
さくら	5	半蔵門	2
西公園	4	特道	2
千代田区	4	徒歩	2
跡	4	下車	2
桜	3	山	2
現	3	所在地	2
現況	2	状況	2
千鳥	3	大通公園	2
周辺	3	町	2
ページ	3	東	2
福岡県	3	もの	2
駅	3	品	2
4月	3	園	2
明治	3	次	2
遠内	3	見頃	2
公園	3	荒津	2
電話	2	先	2
事	2	改称	1
昭和	2	上旬	1
高速	2	道	1
ヤマザクラ	2	樹路樹	1
ソメイヨシノ	2	北の丸公園	1
岡市	2	博多駅	1
東京都	2	城跡	1
観光地	2	都市	1
地	2	完成	1
1日	2	前後	1
問い合わせ	2	建設	1
地下鉄	2	門	1
方	2	公	1
		風	1

【図9】

本発明の一実施例の個別文書解析結果の例

単語	頻度	単語	頻度	単語	頻度
さくら	4	笠備	1	東西	1
千代田区	4	ヤマザクラ	1	東西	1
銀	4	門	1	徒歩	1
緑	3	当時	1	特徴	1
千鳥	3	事	1	美術	1
淀内	3	ソメイヨシノ	1	端	1
緑	3	分	1	状況	1
1日	2	九段下	1	雰囲	1
町	2	若木	1	公園	1
もの	2	高速	1	観光地	1
昭和	2	交通	1	じゅうへん	1
ページ	2	羅	1	曲町	1
4月	2	品	1	問い合わせ	1
周辺	2	すべて	1	学	1
駅	2	見頃	1	前後	1
東京都	2	下京	1	建設	1
半蔵門	2	遊	1	安	1
先	1	とし	1	國立	1
ちどり	1	韓国	1	地下鉄	1
電話	1	東	1	まつり	1
首都	1	完成	1	料	1
ふち	1	所在地	1	九段南	1
シダレザクラ	1	企画部	1	新市	1
皇因	1	羅	1	技術	1
田	1	一時	1	近代	1
広報課	1	街路樹	1	再度	1
10日	1	道路	1		
神社	1	振芸	1		
14日	1	北の丸公園	1		
地	1	次	1		

【図10】

本発明の一実施例の対象文書集合全体特徴の例

単語	スコア	単語	スコア	単語	スコア
植	456.643705463183	見頃	124.53919239905	上旬	54.025448448309
0	456.643705463183	毎日新聞	124.53919239905	先	50.0004282908523
1	415.130641330166	岱	124.53919239905	福	50.0004282908523
裁	373.61757719715	特産	124.53919239905	.	41.5130641330166
さくら	332.104513064133	山	91.9104164998939	勢	41.5130641330166
3	290.591448931116	市	91.9104164998939	照	41.5130641330166
5	290.591448931116	周辺	91.4804036214429	ノリ	41.5130641330166
公園	211.784913941575	昭和	91.4894036214429	樹	41.5130641330166
ヶ	211.784913941575	ヤマザクラ	83.0261282660332	姻	41.5130641330166
4月	207.565320665083	大濠公園	83.0261282660332	立願	41.5130641330166
4	207.565320665083	熊本県	83.0261282660332	淹	41.5130641330166
西公園	166.052256532066	高速	83.0261282660332	1955	41.5130641330166
8	166.052256532066	荒津	83.0261282660332	首都	41.5130641330166
9	166.052256532066	1日	83.0261282660332	名駅	41.5130641330166
玉名	166.052256532066	地下鉄	83.0261282660332	勝	41.5130641330166
線	166.052256532066	2	83.0261282660332	田	41.5130641330166
寺	166.052256532066	6	83.0261282660332	山一	41.5130641330166
緑	166.052256532066	半蔵門	83.0261282660332	1960	41.5130641330166
千代田区	166.052256532066	自動車道	83.0261282660332	坊	41.5130641330166
下車	124.53919239905	岡市	83.0261282660332	黒田	41.5130641330166
千鳥	124.53919239905	蛇	83.0261282660332	公	41.5130641330166
駅	124.53919239905	町	83.0261282660332	麹町	41.5130641330166
ソメイヨシノ	124.53919239905	折	83.0261282660332	人形	41.5130641330166
7	124.53919239905	サトザクラ	83.0261282660332	1900	41.5130641330166
徒歩	124.53919239905	事	72.3427864813708	門	41.5130641330166
淵	124.53919239905	問い合わせ	72.3427864813708	前後	41.5130641330166
明治	124.53919239905	状況	72.3427864813708	端	41.5130641330166
道内	124.53919239905	所在地	72.3427864813708	完成	41.5130641330166
観光地	124.53919239905	群	72.3427864813708	092-741-2004	41.5130641330166
福岡県	124.53919239905	交通	59.305729403164	玉	41.5130641330166

【図11】

本発明の一実施例の個別文書特徴の例

単語	スコア
緑	8.85496183206107
千代田区	8.85496183206107
道内	6.6412213740458
千鳥	6.6412213740458
淵	6.6412213740458
町	4.42748091603054
もの	4.42748091603054
東京都	4.42748091603054
1日	4.42748091603054
半蔵門	4.42748091603054
線	4.08660708744492
駅	2.39380150814102
堀	2.21374045801527
東西	2.21374045801527
1955	2.21374045801527
若木	2.21374045801527
館	2.21374045801527
国立	2.21374045801527
再度	2.21374045801527
完成	2.21374045801527
科学	2.21374045801527
広報課	2.21374045801527
ふち	2.21374045801527
近代	2.21374045801527
皇居	2.21374045801527
シダレザクラ	2.21374045801527
10日	2.21374045801527
首都	2.21374045801527
14日	2.21374045801527
すべて	2.21374045801527

単語	スコア
田	2.21374045801527
美術館	2.21374045801527
新宿	2.21374045801527
建設	2.21374045801527
しゅうへん	2.21374045801527
端	2.21374045801527
神社	2.21374045801527
まつり	2.21374045801527
整備	2.21374045801527
九段南	2.21374045801527
一時	2.21374045801527
麹町	2.21374045801527
前後	2.21374045801527
門	2.21374045801527
九段下	2.21374045801527
当時	2.21374045801527
北の丸公園	2.21374045801527
御苑	2.21374045801527
三	2.21374045801527
技術	2.21374045801527
1979	2.21374045801527
街路樹	2.21374045801527
撤去	2.21374045801527
ちどり	2.21374045801527
03-3264-0151	2.21374045801527
建国	2.21374045801527
道	2.21374045801527
道路	2.21374045801527
際	2.21374045801527
企画部	2.21374045801527

単語	スコア
ケ	2.0025074811164
さくら	1.6914042887972
昭和	1.33500498741094
周辺	0.69967103588392
ページ	0.69967103588392
安	0.667502493705468
東	0.596120943710052
ヤマザクラ	0.596120943710052
地下鉄	0.596120943710052
サトザクラ	0.596120943710052
高速	0.596120943710052
次	0.596120943710052
4月	0.271930638610556
品	0.0117976692737065
ソメイヨシノ	0.0117976692737065
事	0.0117976692737065
先	0.0117976692737065
下車	0.0117976692737065
観光地	0.0117976692737065
所在地	0.0117976692737065
交通	0.0117976692737065
電話	0.0117976692737065
見頃	0.0117976692737065
毎日新聞	0.0117976692737065
問い合わせ	0.0117976692737065
状況	0.0117976692737065
徒歩	0.0117976692737065
特産	0.0117976692737065
7	0.00959380705870577

【図12】

本発明の一実施例の特徴情報の例

単語	スコア	単語	スコア	単語	スコア
千代田区	1470.38639371906	神社	91.8991496074414	一時	43.7624975239632
緑	1470.38639371906	14日	91.8991496074414	美術館	43.7624975239832
渾	827.092346466973	近代	91.8991496074414	道	27.7524977338144
道内	827.092346466973	前後	91.8991496074414	際	27.7524977338144
千鳥	827.092346466973	緋国	91.8991496074414	東	23.5688249194848
縁	678.590328430163	門	91.8991496074414	もの	20.8193465069299
さくら	561.722997725584	新宿	91.8991496074414	安	20.5940958834634
ヶ	424.100874555597	当時	91.8991496074414	すべて	15.008489529514
町	367.506598429766	九段南	91.8991496074414	三	15.008489529514
1日	367.506598429766	まつり	91.8991496074414	次	12.489114897965
半蔵門	367.506598429766	九段下	91.8991496074414	科学	4.1374457224585
駅	268.122106587511	建設	91.8991496074414	ページ	1.48577944532577
東京都	136.604656005689	完成	91.8991496074414	見頃	1.46927220353849
昭和	122.138810129879	麹町	91.8991496074414	ソメイヨシノ	1.46927220353849
1955	91.8991496074414	御苑	91.8991496074414	徒歩	1.46927220353849
広報課	91.8991496074414	03-3264-0151	91.8991496074414	下車	1.46927220353849
しゅうへん	91.8991496074414	端	91.8991496074414	観光地	1.46927220353849
堀	91.8991496074414	北の丸公園	91.8991496074414	毎日新聞	1.46927220353849
東西	91.8991496074414	撤去	91.8991496074414	特産	1.46927220353849
ちどり	91.8991496074414	1979	91.8991496074414	7	1.19480498312352
皇居	91.8991496074414	道路	91.8991496074414	問い合わせ	0.853476269245578
シダレザクラ	91.8991496074414	街路樹	91.8991496074414	状況	0.853476269245578
国立	91.8991496074414	周辺	64.012485804217	所在地	0.853476269245578
首都	91.8991496074414	4月	55.4433702037289	事	0.853476269245578
10日	91.8991496074414	ヤマザクラ	49.4936139345395	交通	0.69866938153446
若木	91.8991496074414	地下鉄	49.4936139345395	技術	0.654223559029463
ふち	91.8991496074414	高速	49.4936139345395	先	0.589888516519153
田	91.8991496074414	サトザクラ	49.4936139345395	品	0.350883341438880
再度	91.8991496074414	館	43.7624975239632	電話	0.286798125124254
企画部	91.8991496074414	整備	43.7624975239632	5	-8.92323790311528

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

BLACK BORDERS

IMAGE CUT OFF AT TOP, BOTTOM OR SIDES

FADED TEXT OR DRAWING

BLURRED OR ILLEGIBLE TEXT OR DRAWING

SKEWED/SLANTED IMAGES

COLOR OR BLACK AND WHITE PHOTOGRAPHS

GRAY SCALE DOCUMENTS

LINES OR MARKS ON ORIGINAL DOCUMENT

REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY

OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.